Information Sciences 339 (2016) 158–174

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Verifying social network models of Wikipedia knowledge community

Michał Jankowski-Lorek[a], Szymon Jaroszewicz[b,c], Łukasz Ostrowski[d], Adam Wierzbicki[e,*]

[a] Polish-Japanese Academy of Information Technology, Koszykowa 86, 02–008 Warsaw, Poland
[b] National Institute of Telecommunications, Szachowa 1, 04–894 Warsaw, Poland
[c] Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01–248 Warsaw, Poland
[d] Institute of Sociology, Warsaw University, Karowa 18, 00–927 Warsaw, Poland
[e] Faculty of Computer Science and Management, Wroclaw University of Technology, 5 Łukasiewicza St., 50–371 Wroclaw, Poland

ARTICLE INFO

Article history:
Received 17 April 2014
Revised 10 September 2015
Accepted 18 December 2015
Available online 6 January 2016

Keywords:
Wikipedia
Social network
Verification
Behavioral network
Survey
Social dimension

ABSTRACT

The Wikipedia project has created one of the largest and best-known open knowledge communities. This community is a model for several similar efforts, both public and commercial, and even for the knowledge economy of the future e-society. For these reasons, issues of quality, social processes, and motivation within the Wikipedia knowledge community have attracted attention of researchers. Research has often used Social Network Analysis applied to networks created based on behavioral data available from the edit history of the Wikipedia.

This paper asks the following question: are the popular assumptions about the social interpretations of networks created from the edit history valid? We verify commonly assumed interpretations of four types of networks created from discussions on Wikipedia talk pages, co-edits and reverts in Wikipedia articles, and edits of articles in various topics, by comparing these networks with results from a survey of editors of the Polish Wikipedia community. The results indicate that while the behavioral networks are strongly related to the declarations of respondents, only in one case of the network created from talk pages and interpreted as acquaintance we can observe a near equivalence. The article next describes improved definitions of behavioral indicators obtained through machine learning. The improved networks are much closer to their declarative counterparts.

The main contribution of the article is a validated model of an acquaintance network among Wikipedia editors that can be derived from behavioral data and validly interpreted as acquaintance. Other contributions are improved versions of behavioral networks based on editing behavior and discussion history on the Wikipedia.

* Corresponding author. Tel.: +48 225844500.
E-mail addresses: m.jankowski@pjwstk.edu.pl (M. Jankowski-Lorek), s.jaroszewicz@ipipan.waw.pl (S. Jaroszewicz), lukasz.ostrowski@uw.edu.pl (Ł. Ostrowski), adamw@pjwstk.edu.pl (A. Wierzbicki).

http://dx.doi.org/10.1016/j.ins.2015.12.015

1. Introduction

Wikipedia is one of the most popular websites on the Internet, ranking sixth globally among all websites on Alexa and having an estimated 365 million readers worldwide. A collaborative effort to organize and present human knowledge similarly to traditional encyclopedias, its most distinctive feature is that anyone can become an editor. This has led to the sustained growth of Wikipedia [24], but also to possible future scalability problems. Nowadays, in the Web 2.0 era, there are many sites where user-contributed content plays a major role. Many other public wiki sites may face similar problems. Due to Wikipedia's openness and lack of centralized supervision, authors have to overcome problems unknown in the editing of traditional encyclopedias [25].

The most notable example is vandalism, which consists mostly of the deliberate deletion of content or the addition of false or irrelevant information. The results of vandalizing Wikipedia may have serious consequences for users, especially when a biographical article is affected. The global impact of this kind of damage, while rather low, is rising [22]. Even though anti-vandalism bots created to automatically prevent this kind of damage do a good job, there is always a need for human reviewers.

Another problem connected with the lack of central supervision arises when editors have different points of view, which may result in an editing war in which two or more contributors or groups try to enforce their versions of an article. This violates one of the key Wikipedia rules mandating contributors to maintain a neutral point of view. Viégas et al. [29] noted that not only controversial articles are threatened by edit wars.

These problems are caused by human factors and cannot be resolved by technology alone. Rather than relying on technology, Wikipedia has created a community of editors who work collectively (in accordance with social norms established by the community) to maintain and improve the sites quality. Since this work involves mostly the sharing, creating, teaching and learning of knowledge, the community of Wikipedia editors is a knowledge community.

The Wikipedia knowledge community can be considered the most developed and mature Web 2.0 community and is, therefore, the object of intense study conducted by researchers wishing to learn how to improve the designs of socially-centred Web platforms. The Wikipedia knowledge community is one of the best examples of how collective intelligence can be achieved on the Web [16]. Wikipedia can also be seen as a model of collaborative work for future economies [31]. For all these reasons, research concerning the Wikipedia knowledge community is of great interest.

Social networks have been the most popular tool for developing models of the Wikipedia knowledge community. Diverse approaches have been used in order to capture various social phenomena, such as editor collaboration[17], conflict[7], mutual trust or respect, and more. This article attempts to give a broad overview, as well as a synthesis, of these approaches. The proposed synthesis uses a Multidimensional (Multilayered) Behavioral Social Network (MBSN) [11]; specifically, we use an MBSN model of the Wikipedia knowledge community (in Section 3) consisting of four dimensions (Discussions, Co-edits, Reverts , and Topics). The proposed model is composed of dimensions that have been widely used in related research. These dimensions, derived from the edit history of Wikipedia, are used to build a behavioral model of the social network.

The proposed model usually incorporates a social interpretation which constitutes the basis of many applications. These interpretations are discussed in detail in Section 3.6. The main contribution of this article is an attempt to verify the proposed social interpretations of the MBSN model of the Wikipedia knowledge community. This verification was conducted by comparing the behavioral social networks that form the MBSN with declarative social networks created on the basis of surveys.*We concluded that the interpretation of a behavioral social network using a social concept is valid only if it matches the declarative social network that concerns the same social concept.* In other cases, the behavioral social network may contribute to an understanding of the social concept (especially if it shows a strong correlation with the declarative network), but cannot be interpreted directly using that concept. For example, if a behavioral social network created from histories of edit reverts among editors is to be interpreted as a conflict among editors, we assume that a survey asking about conflict would result in a declarative social network that would closely match the behavioral one.

The initial results of our validation, given in Section 4, indicate that the four dimensions that have been widely used in research on the Wikipedia knowledge community are indeed correlated with their respective social interpretations.*However, only the Discussions dimension (created from article talk pages; see Section 3.2) can be seen as an approximate equivalent to the social interpretation of acquaintance.*

Our initial result could be the result of simplistic definitions of the behavioral social networks. In the example of a behavioral network constructed by reverts, perhaps only the most recent reverts should count in order to create an indication of conflict. In order to improve our behavioral social networks, in Section 5 we use machine learning techniques to obtain more complex, synthetic behavioral indices which are more closely related to their respective social interpretations.*However, the obtained results still do not support the social interpretations proposed in the literature for three of the behavioral social networks: Co-edits, Reverts, and Topics.* Again, only the declarative acquaintance network can be approximated by our model, which uses the Discussions network along with other variables derived from Wikipedias edit history.

The paper is organized as follows: Section 2 discusses related work, with special attention given to work using the social interpretations discussed in this paper. Section 3 introduces the MBSN model and states five research hypotheses (based on reviews of related work) regarding possible social interpretations of the MBSN dimensions. Section 4 introduces our survey method for constructing the declarative social networks needed to validate our hypotheses and shows the initial results of testing the validity of the hypotheses. Section 5 describes our attempts to improve the behavioral networks using Machine Learning to match the survey data as closely as possible, in order to preclude the possibility that the mismatch of

three MBSN dimensions with declarative networks, as reported in Section 4, was due to overly simplistic definitions of the former.

Finally, Section 6 states our conclusions. In brief, our results do not preclude the use of the discussed behavioral social networks for many applications, for example for recommendation systems. Nevertheless, applications that rely directly on social interpretation for example, interpreting one of the networks as trust, or another as expertise should be treated with caution and evaluated independently.

2. Related work

Collaboration on Wikipedia has important social aspects that have been the subject of intense study. Research has focused not only on the methods and efficacy of work coordination and on the effectiveness of work distribution, but also on negative aspects of collaboration such as conflict and edit wars, or on methods of conflict resolution, moderation, etc.

One of the early papers on Wikipedia collaboration [12] comprehensively studies the effect of conflict on collaboration on three levels: global, article and user. On the global level, the problem consists of the growing cost of coordination. As Wikipedia grows exponentially in size, the number of new articles decreases. A greater portion of the work force is being focused on maintenance and indirect work. The Conflict Revision Count was proposed in [12] as a good predictor (in machine learning terms) of growing conflict. Based on this, automatic detection and prediction of controversy was attempted in large-scale social collaborative systems. Conflict on the user level was shown by means of network visualization. It was also shown that the Reverts network can be used to cluster users sharing similar points of view.

Research results obtained for Wikipedia have been generalized to other types of collaborative editing using Wiki technology. In [13], the authors examined the correlation of coordination and conflict in wiki production groups. They identified four mechanisms of coordination: intra-article communication, inter-user communication, concentration of workgroup structure and policies and procedures. Next, they identified conflicts and analyzed mechanisms for managing them. A notable finding is that most wiki production groups other than Wikipedia share similar direct coordination mechanisms and problems with conflict management while differing in the use of policies and procedures. Wikipedia has developed a large base of policies and procedures for enforcing them, while most other wiki production groups have not. This is most likely related to the size and age of Wikipedia, as well as its popularity; other wikis are more condensed, with smaller user bases. Recent research also recognizes the role of social norms and context in online social networks, similarly to Wikipedia [4].

Recent research on Wikipedia collaboration has applied advanced statistical methods. In [8], the authors present an examination of Wikipedia articles with varying needs for coordination. Using statistical network analysis methods called p*/exponential random graph models (p*/ERGMs), they analyze the multi-level processes which structure diverse Wikipedia collaborations. This diversity is an intrinsic feature of Wikipedia. For example, it is rather difficult to predict the effects of co-editing an article about current events, especially breaking news such as natural disasters or commercial airline crashes. Traditional network-based approaches, including visualization and descriptive statistics, clearly show patterns of collaboration, especially for typical encyclopedic articles, but give no insight about their statistical significance, effectively limiting their application in the validation of hypotheses. Demands for coordination regarding a particular article show a correlation with the number of authors. However, the edit histories of disputed or breaking articles show a fair amount of more complex interactions between authors and their attributes. Surprisingly, the actual number of experienced Wikipedians editing breaking news is significantly smaller than in a random model.

New research also attempts to capture the qualitative aspects of collaboration through a more in-depth analysis of the semantics of collaborative actions. Paper [21] contains an analysis of 58 talk pages in the following categories: articles with most editors, most views, and controversial articles. The types of comments in the discussions usually vary by article type, but include: requests/suggestions for editing coordination, references to Wikipedia policies, references to outside sources, and references to deleted (or reverted) content. A surprising finding is that most controversial pages (at least compared to the other two groups) generate comparatively little significant discussion on the talk pages.

Earlier, we conducted some research to find an approach to the problem of evaluating teams of contributors (instead of single authors) by means of multidimensional implicit social networks [9,26–28]. We presented the idea of creating such a network based on the edit histories of certain pages and collaboration patterns between contributors. Our social network is based on Wikipedias entire edit history taken as a whole, and therefore is a summary of all recorded author interactions. This social network can be used to assess the quality of a team of authors and consequently, to recommend good teams. It can also be used by Wikipedia authors and editors as an additional tool enabling them to improve collaboration, as it expresses each authors social environment and can be navigated to discover new projects in which an author can participate, or to recommend new collaborators.

Behavioral social networks are the subject of numerous research efforts in other domains, for example in the study of Peer-to-Peer networks [15], grids [23], in the study of reputation systems for e-commerce [10,18,19] and recommender systems [2]. However, little research is concerned with the comparison of behavioral and declarative social networks.

3. Multi-Dimensional Behavioral Social Network (MBSN)

The study of multidimensional behavioral Wikipedia social networks [26,28] is the subject of our ongoing research, which attempts to model the community of Wikipedia contributors while emphasizing the aspect of teamwork. The research tool

we have used is a social network analysis performed on the behavioral social network based on Wikipedias edit history. To create this dataset, we analysed the entire edit history of the Polish edition of Wikipedia since its inception in 2001. Our goal was to find the actual social relationships between authors, such as trust, criticism, acquaintance and common interests. In this section we describe the Multidimensional Behavioral Social Network (MBSN) used in our work and the approach we used to construct it.

3.1. Pre-processing

The main obstacle was the amount of data present in the edit history and the difficulty of performing operations on this data. In the case of the Polish Wikipedia, the edit history constitutes over 220 GB of text. First, we needed a way to concisely represent the articles text and information on its authorship. We assumed a single word as the basic unit of content. We processed each version of a particular article in order to observe the changes that were made and assigned an author to each word. Thus, the first version consisted of words written by the creator of the page, while subsequent revisions contained all words of the text at a particular time along with their respective authors.

Between every pair of subsequent revisions there may be four kinds of actions: adding a word, deleting a word, moving a word from one place to another, and changing a word. Adding is simply placing a new word in the text (whose author is the author of the revision in which it first appeared). Deleting is simply removing a word from the text. Moving is removing a certain portion of text from one place and placing that text, in exactly the same sequence, in another place. Changing is an operation of replacing one word with another (including, for instance, spelling corrections). We needed to distinguish moving from deleting followed by adding in order to preserve authorship information. A threshold used to avoid identifying the moving of single words or common phrases as moving text written by a previous author functions by identifying how many consecutive words were moved; if the result is below this threshold, then the whole operation is considered a deletion followed by an addition by a new author. A replacement of a single word is also considered as a deletion followed by an addition.

The MBSN is a set of graphs consisting of nodes [11], each representing one Wikipedia contributor (some graphs may also contain other nodes, such as Wikipedia categories) and edges, each representing one kind of relationship between them. The specific weight of each edge is represented by a numeric value. We have defined four dimensions (networks) of relationships between authors: Discussions, Co-edits, Reverts, and Knowledge (interests). This network is completely behavioral, meaning that it contains no declared information about social relationships and is based solely on the edit history.

3.2. Discussion

To calculate the edge strength in the Discussions network, we looked at the articles and users talk pages. The measure is proportional to the amount of text added by one author next (that is, in response) to text written by the other author. Activity on talk pages has been used in the literature to evaluate the degree of collaboration between editors. It has been hypothesized that the Discussions network can be interpreted as the social relationship of acquaintance.

The strength of the discussion edge between authors A_1 and A_2 is computed as the total number of words written by A_1 following a text written by A_2 but no further than *discussion_distance* words away.

The value of the parameter *discussion_distance* was chosen as 20 based on empirical evaluation. A typical case for discussion on talk pages is that at least 20 words are written by each participant. Increasing the value of the parameter would result in ignoring shorter exchanges. As a side effect, after each exchange between authors A_1 and A_2, the value of the strength of the edge between them usually increases by $210 = \sum_{i=1}^{20} i$.

3.3. Co-edits

The main operation that influences edge strength in the Co-edits network is the addition of text in the vicinity of text written by another author. We believe that someone who edits the text of an article has read (reviewed) the surrounding paragraphs. For this reason, we hypothesized that the Co-edits network could be interpreted as the social relationship of trust [32].

Co-edits are defined as the amount of text (number of words) written by one author next to the text of the other author. The exact measure is calculated as follows. Let w_i stand for a word added to article k by author A_1 and w_j for a word previously added to the same article by author A_2. Let $D_k^{i,j}$ be the distance in words between w_i and w_j in the article k (values larger than *distance_cutoff* are ignored). We use the value of 20 for *distance_cutoff* based on empirical evaluation. Then we define the relationship based on all articles that both authors have edited:

$$Coedits(A_1 \to A_2) = \sum_k \sum_{i,j} \left\{ \frac{1}{D_k^{i,j}} : D_k^{i,j} < \text{distance\_cutoff} \right\}.$$

3.4. Reverts

The edge strength in the Reverts network is measured by the number of edits made by one author and reverted by another. This measure enables detection of edit wars, in which two or more authors or groups argue with each other. Revert

operations have frequently been used in the literature to model conflict and to identify edit wars [14]. We hypothesized that the Reverts network could be interpreted as the social relationship of distrust or criticism [30].

Edge strength in the Reverts network is measured by the number of edits made by one author and reverted by another. The strength of an edge in the Reverts network between authors A_1 and A_2 is computed as follows: for each revision R in the edit history we look to see whether there was an identical revision R_0 in the last *max_recent* revisions. Each such pair (R_0, R) increases the value of $Reverts(A_1 \rightarrow A_2)$ by $|\{R_i : author(R_i) = A_2$ and R_i lies between R_0 and $R\}|$. max_recent is a parameter describing how far back in the edit history we should look when trying to match a revert. In our research we used a *max_recent* value of 20.

3.5. Topics

The Topics dimension is a bit different from the others, because the set of nodes is extended by a subset of Wikipedia categories and the edges form a bipartite graph connecting authors to the categories of the articles that they have edited. The strength of the edges is proportional to the number of different articles in a particular category in which the given editor has made at least one edit. Not all categories have been added to the set of nodes: we have attempted to filter out non-topical categories (for example, dates or the "disambiguation" category). We hypothesized that the Topics dimension could be interpreted as a relationship of interest or knowledge of an author in a topical category.

The edge strength in the Topics network *Edits-in-category*($A \rightarrow C$) is defined as the total number of changes of any type made by author A in articles belonging to category C, where only changes exceeding the minimum length of 20 words are counted in order to eliminate nonessential revisions.

3.6. Possible social interpretations

The MBSN can be considered simply as a set of behavioral social networks that can be used for various purposes, such as recommendation. For example, in [20] we described how the four dimensions of this network were linked to the votes cast in the RfA (Request for administrators) procedures. The MBSN may also be a new valuable tool for recommending candidates for admins.

However, we can consider the MBSN as a valid social model of the community of Wikipedia editors only if we can reliably interpret these dimensions as meaningful social relationships. In the literature, this interpretation has usually been assumed, and only partially validated through indirect evidence. For example, if it can be found that the Reverts dimension is related to edit wars, this can be argued as an indirect validation that the Reverts network can be interpreted as a social relationship of conflict or distrust. In [20], we showed that the strength of edges in the Reverts dimension was higher for votes against than for votes in favor of a candidate in the RfA procedure. This finding can also be used as indirect evidence proving the interpretation of the Reverts dimension as conflict or distrust.

The Topics dimension links the authors with article categories. A high topic index value means that the author edited a large number of articles in a given category. But why do authors edit a number of articles in a specific category? Is this meaningful behavior or merely random choice? We have proposed two possible interpretations of this "behavior-based indicator": that authors editing articles in a category either consider themselves experts on or are merely interested in this topic. Even though expertise and interest are not independent, we assume them to be two distinct attitudes towards article categories. For instance, an editor may be very interested in quantum physics but, due to lack of a suitable education, does not consider himself an expert. On the other hand, one may be a professional quantum physicist without being interested in the topic as a Wikipedia editor.

In this paper, we have attempted to directly check the validity of the hypotheses concerning the interpretation of dimensions of the behavioral social network. We used a survey of over 100 Wikipedia editors to obtain declarative data that can be used to validate our behavioral social networks (the survey results included several thousand declared relationships). We tested the validity of the following five hypotheses regarding the current dimensions of the behavioral social network:

Hypothesis 1. The Discussions network can be interpreted as acquaintance among editors

Hypothesis 2. The Co-edits network can be interpreted as trust in the ability of an editor to produce content of good quality

Hypothesis 3. The Reverts network can be interpreted as a state of conflict between editors

Hypothesis 4. The Topics network can be interpreted as the interest of an editor in a topic

Hypothesis 5. The Topics network can be interpreted as an editors expert knowledge about a topic.

4. Survey based verification of MBSN validity

4.1. Behavioral vs. declarative Social Networks

The common meaning of acquaintance, trust or conflict is quite obvious and intuitive, but for application in network analysis a more precise definition, as well as operationalization, is neededone that allows for measurement and empirical research.

Consider, for example, acquaintance. How do we know if two people are acquainted or not? One way to find out is to simply ask them. If they declare "yes, we know each other", we can assume they are acquaintances. We call this a "declaration-based index" of acquaintance.

Another way is to observe how people behave. For example, if they shake hands, we can be quite sure they know each other. In the case of Wikipedia, we cant watch people shaking hands, but we can observe how they communicate via talk pages, e.g., post next to each other. In this case we base our knowledge on the "behavior-based index" of acquaintance.

Both indices have their strengths and weaknesses. The declaration-based index is very straightforward and easy to understand, but depends heavily on a persons memory—usually we have more acquaintances that we can name.

On the other hand, the behavior-based index is totally independent of memory; in Wikipedia edit histories one can detect traces of acquaintance for as long as one year after the last interaction. But this index can be misleading too. For example, small talk may not be enough to constitute a relationship.

While conscious of the strengths and weaknesses of the presented measures, we consider both of them good indices of the social relationship of acquaintance in terms of face validity [1]; in our opinion they adequately depict the meaning of the concept. But can we prove their usefulness by showing some evidence that they are measuring the same thing? In social science this is a question of so-called criterion-related or predictive validity [1], which is tested by studying the ability to predict the value of one index based on the value of another.

In order to test the predictive validity of the MBSN dimensions that are behavioral indices of real social relations, we used the survey method.

4.2. Survey methodology

4.2.1. The questionnaire

To test the predictive validity of the behavior-based indices we asked the respondents the following questions. For testing the Discussions dimension:

Q1. "We would like to know how many Wikipedians you know. Please name everyone that you can remember (use nicknames)."

Q2. "Now, please look at the list of Wikipedians with whom you have edited the same articles. Mark nicknames that you recognize (i.e., you remember that you have seen them before)."

Both questions are declaration-based indices. We shall refer to the first declarative network as unsupported Acquaintances, to the second as supported Acquaintances.

The next question concerned whether one editor contacted another one:

Q3. "Please mark the nicks of editors with whom you have been in contact."

For creation of a declarative Trust network we used the following question:

Q4. "Please mark the nicks of editors characterized, in your opinion, by edits of good quality."

For creation of a declarative Conflict network we used the following question:

Q5. "Please select the nicks of editors with whom you have at any time disagreed or argued.

For testing hypotheses regarding the Topics dimension we used two questions:

Q6. "Please look at the list of article categories. For each category, try to determine your level of competency." [Scale] 0 = "I don't know much about it", 100 = "I know a lot about it".

Q7. "And now, for each category, please determine the extent to which you are interested in that topic." [Scale] 0 = "I'm not interested at all", 100 = "I'm very interested".

We shall refer to the declarative network obtained from Question 6 as the Expertise network, and to the network obtained from Question 7 as the Interest network.

4.2.2. Survey sample

The sample of network relationships to study was obtained in two steps. The first was to select a sample of Wikipedians. The second was to select a sample from the set of possible relationships of each chosen Wikipedian.

Drawing a random sample from the Wikipedia community is a difficult task. Although it is easy to list all the Wikipedians, it is hard to contact respondents within the chosen sample and convince them to answer the questions. Because of the difficulty in obtaining a representative random sample, we had to rely on non-probabilistic sampling techniques. We decided to rely on subjects available through informal relationships and the Polish Wikipedia discussion group, and then to snowball (by asking respondents to invite others) [1].

4.2.3. Sampling, the second step

The second step of the sampling procedure involved selecting a subset of relationships for every respondent who agreed to answer the questionnaire. Each Wikipedian interacts with many others.

An average Wikipedian has 14.5 neighbors in the Discussions network (std dev 62; median 2; min 1; max 1,669); 43.2 in Co-edits; (std dev 252.1; median 6; min 1; max 13,514), and 16.3 in Reverts (std dev 78.7; median 2; min 1; max 2,265). Because of the large range of and variance in the numbers of possible relationships, we had to limit the number of relationships to be asked about.

An average Wikipedian has 14.5 neighbors in the Discussion network (std. dev. 62; median 2; min 1; max 1669), 43.2 in Co-edits; (std. dev. 252.1; median 6; min 1; max 13514), and 16.3 in Reverts (std. dev. 78.7; median 2; min 1; max

2265). Because of the large range of and variance in the numbers of possible relationships, we had to limit the number of relationships to be asked about.

In the first question of the survey the respondent was asked for his or her nickname. Based on the answer, the survey software generated a subset of relationships to study for each individual respondent. The subsets were then calculated separately for the Discussions, Co-edits and Reverts networks. The distribution of relationship strength was expected to fit a power law. To ensure that links of various strength were included, stratified sampling was applied.

First, the respective range of the edge strength was calculated (the difference between the largest and smallest value) separately for each respondent and dimension. The range was then divided into four equal intervals, each defining one stratum. From each stratum two network relationships were picked at random. If there were less than two elements in the stratum, the missing element was taken from a higher stratum. Therefore, every individual network was represented by up to eight elements. Each respondent was asked about a maximum of 24 relationships. The nicknames were displayed in the order they had been chosen for the sample.

A sample of $n = 111$ Polish Wikipedia editors participated in the survey. We asked each editor about his/her relationships with other Wikipedians that we had detected using the MBSN dimensions, so that we could compare the values of behavior-based and declaration-based indices. Respondents answered questions about 818 Discussions, 745 Co-edits, and 419 Reverts relationships.

Sampling for knowledge and interest. To investigate the relationship between the Topics index and expertise/interest, for each respondent we randomly chose up to 10 topics that he/she had previously edited. Next, we asked separately about each selected topic.

4.2.4. The sample and the population

It should be noted that the sample of users who answered the questionnaire is not identical to the general population of Polish Wikipedians. Users with reviewer or administrator status participated much more often than ordinary users. In June 2011, the MBSN of the Polish Wikipedia community included over 90,000 users, compared to 2,349 reviewers (2.6% of users) and 115 administrators (0.1% of users). The sample included 58 reviewers (53%), 28 administrators (25%) and 25 regular users (23%). Therefore, the sample was skewed towards users who are more experienced and involved in the community.

This can be seen as a disadvantage of our study. However, in our research the unit of analysis is not a user, but a single network relationship. We simply would like to know if behavioral network ties match users perceived social ties. Therefore, it was more important to take a sample which was representative in terms of edge strength rather than user status.

One might suppose that the relationship between behavioral and perceived ties depends on the history of the individual relationship (length, frequency, quality etc.) rather than on attributes such as user status. We applied post-stratification survey weights to match the distributions of edge strengths in the sample and in the population (described below).

It is also worth mentioning that recommendation algorithms based on the MBSN or other behavioral data would be most useful for the most active Wikipedia editors, such as reviewers or admins. Thus we believe that our research provides useful evidence on the validity of the proposed indicators despite certain sampling inadequacies.

To correct the difference between the sample and population structure, each sample unit was assigned a weight proportional to the population frequency. A separate set of weights was prepared for each network dimension. To calculate the weights for a given dimension, the full range of a network tie strength was first divided into intervals. The thresholds of the intervals were based on the values of the edge strengths from the sample. Each consecutive boundary was set midway between two adjacent values of tie strength. Next, the population frequency of the ties in each interval was calculated using the Wikipedia database dump. In the last step, weights were assigned according to the following expression:

$$w_t = \frac{n}{N}\frac{N_t}{n_t}$$

Where: w_t – weight assigned to each tie of strength t, N – number of ties in the population, n – number of ties in the sample, n_t – number of ties of strength t in the sample, N_t – number of ties of strength t in a particular interval in the population.

4.3. Analysis

4.3.1. Discussion as acquaintance

As a first step, we assessed the value of our behavior-based index for predicting an acquaintance relationship operationalized with our declaration-based index. We learned that 45% of editors identified as acquaintances by the behavior-based index are recognized by the respondent, and that 6% of them are recalled.

We can improve on this result by increasing the cutoff point for the strength of the behavior index: edges with low strength are dropped and no longer treated as indicators of acquaintance relationships. We learned that prediction validity varies along with the behavior-based indexÃ¥ cutoff point. For recognition, it ranges from 45% of recognized editors for very low cutoff values to 96% of recognition for very high values. For recall, prediction validity increases from 6 to 53% (Fig. 1). This result is very encouraging, and shows that it is indeed possible to use behavioral social networks as indicators of declared real social relationships.

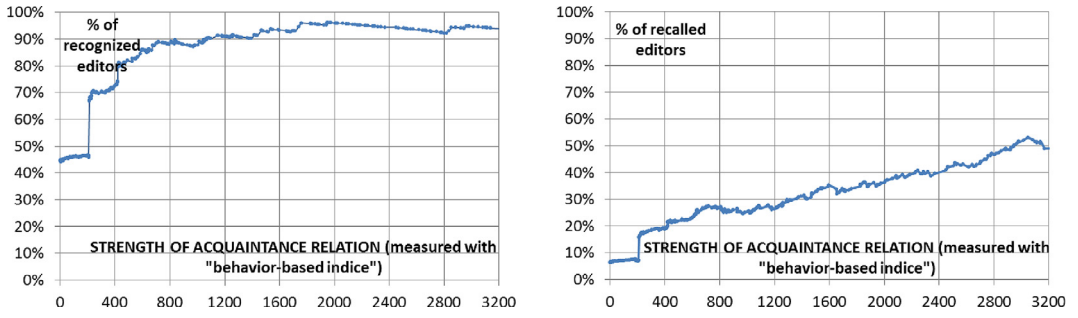

Fig. 1. Predicting recognition and recall of acquaintances with Discussion network indicator.

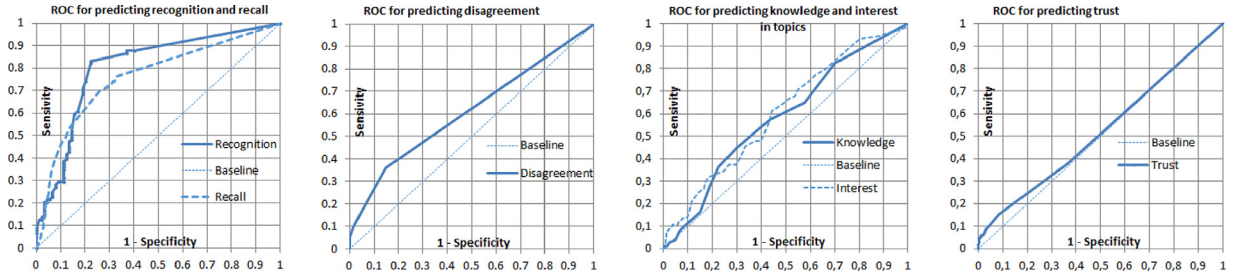

Fig. 2. ROC charts for predictions based on behavioral social network.

The stepwise shape of the plot in Fig. 1 is explained by the typical increase in strength of the edge in the Discussion network of 210 (see Section 3.2).

Achieving a satisfactory level of recall of declarative acquaintance with the behavior-based index is not enough to declare its validity. For example, if all editors were declared acquaintances, the recall of behavior-based index would be 100%, even if only a few declared acquaintances were discovered within the index. To obtain more meaningful results, we need to include information concerning the number of false positives. This can be achieved using so-called ROC curves [6], which visualize the tradeoff between false positives and negatives in a single chart. ROC curves are the most popular method of evaluating classification models in Machine Learning and their use enabled us to easily compare (see Section 5.1) operationalizations based on single, predefined dimensions with more complex operationalizations based on Machine Learning methods (see Section 5).

The ROC curves for predicting declaration-based acquaintances using our behavior-based measurement are shown in Fig. 2 (first chart). In order to obtain a ROC curve, all examples are sorted based on their *score*, which here is the value of the behavior-based index. Each value of the score is used as a cutoff point at which two values are computed. One value is *sensitivity*: in our case, this is the percentage of all declared acquaintances correctly identified using the behavior-based index at a given cutoff. The other is *specificity*: in our case, the percentage of all non-acquaintances incorrectly declared as acquaintances. Every cutoff threshold corresponds with a point on the ROC curve; the y-coordinate of the point is the sensitivity corresponding to the threshold, while the x-coordinate corresponds to $1 - specificity$.

ROC curves enable us to visualize the tradeoff between a models sensitivity and specificity. For example, in the recognition plot in Fig. 2 we can see that we are able to choose, lets say, 50% of our acquaintances, but we have to pay for that by incorrectly identifying about 30% of our non-acquaintances as acquaintances. If we agree to identify only 20% of our acquaintances, the penalty is lower: only 10% of non-acquaintances are incorrectly classified. The diagonal line in the plot corresponds to a model making random predictions and is used as a baseline (i.e., total absence of any relationship). The higher the curve is above the diagonal, the better the model. The ROC curve of a perfect predictor passes through the upper left corner of the plot. In our case this would correspond to the *equivalence* of behavior-based and declarative indices. More details on ROC curves can be found in [6].

By looking at the leftmost ROC curve in Fig. 2, we can see that the behavior-based index for recognition is strongly related to its declarative counterpart.

4.3.2. Co-edits as trust

In order to check the validity the Co-edits network, we applied the same procedure as for the Discussions network.

In order to check the validity of the hypothesis that the Co-edits dimension may be interpreted as trust in an editors competence, we needed a declaration-based indicator of trust between authors. To obtain it, we asked our respondents to: *Please mark the nicks of editors characterized, in your opinion, by edits of good quality (Q4)*. Respondents were presented with a list of Wikipedians linked to him/her via a Co-edit relationship and which he/she had already recalled or recognized.

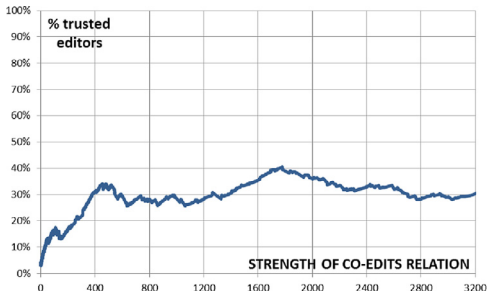

Fig. 3. Predicting "trust in competence" with Co-edits indicator.

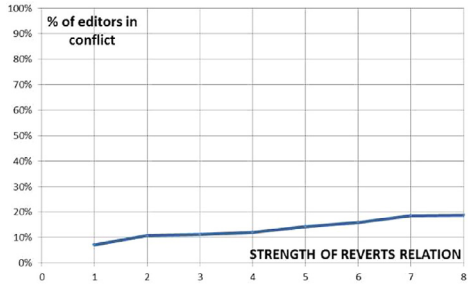

Fig. 4. Predicting "antagonism / disagreement" with Reverts indicator.

(We assumed that one has to know somebody to trust him/her.) In other words, the declarative trust relationship is a subset of the acquaintance relationship. A total number of 745 relationship detected by the Co-edit indicator were evaluated.

As shown in Fig. 3, we learned that only 3% of editors linked to respondents via a Co-edit relationship were recognized and trusted as authors of good-quality edits. Raising the cutoff point for Co-edit relationships improves this result, starting from 3% for Co-edit indicator strength equal to 0.05 and up to 40% for a strength around 1750.

The results show that the behavioral Co-edits network is correlated with declarations of trust in an editors competence. Also, the achieved prediction quality, as visualized by the ROC curve in Fig. 2, is higher than that of random choice, but not significantly so. It is thus again difficult to talk about the equivalence of the indices.

Although the increase in prediction quality is considerable, the overall maximum result of 40% declared trusted editors seems insufficiently high to claim that the Co-edits network is a good equivalent of trust in competence. What explanation can be offered for this? A closer enquiry reveals that most of the editors linked via a Co-edit relationship do not know each other. Therefore, the most probable explanation is that co-editing is usually anonymous: editors do not know who they are working with. They pay no attention to the authorship of the text in the vicinity of their contribution. Moreover, it is not easy to determine authorship with the default Wikipedia interface. In order to do so, one has to dig deep into an articles history and search for chosen text, edition after edition. This obstacle breaks the link between text trustworthiness and personal trust. One can speculate that an user interface with easier access to authorship information would be a better environment for building personal reputations and trust between authors. Future work on trust in competence behavioral indicators therefore should focus on networks of acquaintances and conscious and deliberate cooperation.

4.3.3. Reverts as conflict

As a declaration-based indicator of antagonism, criticism, and conflict we used the survey question:*Please select the nicks of editors with whom you have at any time disagreed with or argued (Q5)*. As previously, the respondent was presented with a list of Wikipedians whom he/she had already recalled or recognized and who had been linked to him/her with a Reverts relationship. (As in the case of Co-edits, an assumption is made here that one has to at least recognize somebody to be in conflict with him/her.) A total number of 419 relationships detected with the Reverts indicator were evaluated with this question.

We discovered (Fig. 4) that only 7% of Reverts relationships are manifestations of a conscious disagreement, argument or conflict. Raising the Reverts indicator cutoff point improves this result to 20%. Clearly there is a strong correlation between behavioral and declarative indices of distrust; the ROC curve in Fig. 2 shows that the behavioral index is a strong predictor, much better than random selection. However, contrary to our expectations, the Reverts network relationship proved to be highly nonequivalent to a conscious, social relationship; only 20% of such declarative relationships can be identified. We offer a few explanations for this.

The first issue is the reason a given contribution was reverted. On one hand, the reason may have been substantive: the information provided was not true or broke the rules of notability or neutrality of point of view. On the other hand, it may have been editorial: vandalism, poor wording, poor formatting, poor quality. It is reasonable to suppose that only reverts

made for substantive reasons can create a personal relationship of disagreement and antagonism. It is possible that authors making editorial reverts do not perceive their actions as manifestations of an argument or disagreement.

A second possible explanation concerns the degree of anonymity of reverts. In Wikipedias user interface it is hard to revert a contribution without seeing who the author is. However, to see is not always to notice. It is likely that most reverts are made by administrators and users working on so-called flagged revisions (checking contributions made by inexperienced users). These Wikipedia shepherds routinely make many reverts a day, usually without being conscious of an important dispute or argument. It may be that in most cases they do not even notice who is being reverted.

Another factor is the amount of time that has passed since the revert. Declaration-based indicators depend on memory. We can suppose that even if an author noticed whose contribution he/she rejected, the memory of this incident would gradually fade. A behavior-based index, on the contrary, is independent of memory. Therefore, many real arguments correctly indicated through a revert, measured just after a relationship bond was established, might be simply forgotten.

Yet another issue consists of the meaning of terms argument and disagreement. Perhaps for many people these terms imply a bilateral relationship, an interaction based on the exchange of conflicting opinions. In the common understanding of the term, reverting may be too one-sided an action to be defined as an argument. Moreover, having ones own contribution reverted triggers much more emotion than reverting the contribution of another. Thus it seems that bilateral reverts or being reverted might serve better as an behavioral indicator of argument/disagreement. Finally, the method of calculating the revert network indicator might matter as well. Our Revert relationship linked one author to another not only when a contribution had been rejected immediately, but also when an author had just restored an older version of an article. In the latter case, the author becomes linked to all of the authors that contributed to the restored version of the article.

4.3.4. Topics as expertise or interest

The validity of Hypotheses 4 and 5 was tested with two different declaration-based indicators. For the creation of an declarative Expertise network we used the survey question:*Please look at the list of article categories. For each category, try to determine how much you know about it [Scale 0100] (Q6)*. For the creation of an declarative Interest network we used the question*For each category please determine the extent to which you are interested in that topic [Scale 0100] (Q7)*. Respondents were able to express their opinion in a simple and intuitive way, using a slider. We collected $n = 1550$ responses about respondents attitudes concerning expertise and $n = 1534$ about their attitudes concerning interest.

A surprising finding was that, in 56% of Topics relationships, respondents declared no knowledge at all; and, in the case of 62% of relationships, no interest. We suppose that in some cases 0 on a 0100 scale means that respondent did not really answer the question. Nevertheless, the apparent irrelevance of some categories to authors knowledge and interest is puzzling.

An even more puzzling result is that average interest and expertise do not increase monotonically for high topic index values. Average interest and expertise increase with the number of edits in a category up to a certain level and start to drop above that level. Respondents showed little interest and expertise in categories that were edited only once (average 910 on a scale of 0100). They declared increasingly more interest and expertise for categories with up to 1120 edits and tended to declare less interest and expertise for categories edited very often.

The relationship between Expertise or Interest and Topics index value can be approximated with a linear model for low values of the behavior-based index. However, the relationship does not fit a linear model for higher values. We found the best linear fit for values ranging from 1 to 18. The fit decreases for higher index values.

In the case of declared expertise, for low values of the Topics index (ranging from 1 to 18), the relationship can be modeled with a linear equation: declared expertise $= 8.7 + 2.8$(topics index) with a weak fit ($R^2 = 0.083$). Interest is very similar here to expertise. A weak linear relationship can be found for low values of the Topics index. This can be modeled with the equation: declared interest $= 10.5 + 2.5$(topics index) with a slightly worse fit ($R^2 = 0.061$).

There may be several reasons for such a weak relationship between Topics index values and declared interest and expertise. First, some topics are artificial categories, abstract lists of objects that have very little in common, e.g., 1853 births. Second, many edits made by reviewers and administrators are editorial: NPOV (neutral point of view) violations, wording, spelling mistakes, etc. One need not be interested in or an expert on a topic to do editorial work. Third, interest and expertise are often focused on a lower level. The author may have expertise in one subject, but not in the whole category; for instance, one may claim to be an expert on Van Gogh but not on Dutch Painters in general. It might be expected that the more general the topic is, the lower the level of declared interest and expertise. So one may claim to be an expert on Van Gogh and declare some expertise in Dutch post-impressionist painters, but not deem oneself to be an expert on Dutch painters at all.

5. Construction of behavioral indexes using machine learning methods

In the previous section, we carried out a verification of commonly assumed hypotheses regarding social interpretations of MBSN dimensions by evaluating their predictive validity for declarative indices of real social relationships. This approach has shown that, although all dimensions are strong predictors of their respective declarative interpretations, only one of the MBSN dimensions, the Discussions dimension, offers sufficient predictive accuracy to be considered a behavioral equivalent. In an attempt to improve the MBSN network, we decided to use a machine classification approach that would enable us to learn how best to construct new MBSN dimensions that could predict the declarative indices of Acquaintance, Trust, Conflict, Interest and Expertise.

Table 1
Categories of input variables for machine learning models.

Variable category	Based on	Description
Cat_similarity	Edits history	Number of categories in which both editors (SRC, DST) contributed edits.
Discussion	Talk pages	One of MBSN dimensions computed based on discussion length.
Reverts	Edits history	One of MBSN dimensions computed based on reverts length.
Cat_diversity	Edits history	Number of different categories in which editors (SRC , DST) contributed edits.
Coedits	Edits history	One of MBSN dimensions computed based on coedits length.
Days_since_last_ < dimension >	Edits history / Talk pages	Number of days since last SRC and DST perform edit/revert/discussion in one article
Avg_days_between_ < dimension >	Edits history / Talk pages	Average number of days between SRC and DST edits/reverts/discussions in one article
Days_since_first_ < dimension >	Edits history / Talk pages	Number of days since first SRC and DST perform edit/revert/discussion in one article
Reverted_ < dimension >	Edits history / Talk pages	Social network dimensions computed in reverted direction DST -> SRC
Encounter_ < dimension >	Edits history/ Talk pages	How many times SRC and DST performed edit/revert/discussion in same article.

Table 2
Target variables of machine learning models.

Target Name	Description
Expert	Indicator if SRC chose DST as a trusted editor of high quality (Question 4)
Contact	Indicator if SRC confirmed that was contacting DST
Argue	Indicator if SRC confirmed that was arguing with DST (Question 5)
Spontaneous	Indicator if SRC provided DST as known Wikipedia user without any hint
Acquaintance	Indicator if SRC checked DST as known Wikipedia user on presented list

The proposed approach uses much more of the information available from edit history than the first simple dimensions of the MBSN. In particular, the machine classification models were able to use information about elapsed time between events. Also, the models were able to use reversed relationships; for the Reverts dimension, we initially hypothesized that the dimension was in the wrong direction for predicting conflict (i.e., that the editor whose edit had been reverted was more likely to feel antagonism towards the reverting editor). Such hypotheses could be automatically checked by the machine classification approach for all survey questions. Moreover, the machine classifiers could use all available information for predicting each survey question, for example, using the Reverts dimension to predict the questions about trust, or acquaintance.

The main goal of this section was to improve the verification of the hypotheses described in Section 4. The negative validation of Hypotheses 25 could be due to simplistic definitions for the MBSN dimensions. If the improvements proposed in this section did not lead to a much better agreement of the MBSN dimensions with declarative indices (such as observed for the Discussions dimension), we would conclude that Hypotheses 25 had been negatively verified.

5.1. Machine learning methods used

To build our classifier, we created an input dataset, with additional variables computed by transforming other information extracted during the creation of the MBSN. Each record represents a single relationship between two users and is characterized by attributes related to them:

- SRC – a user who responded in our survey
- DST – a user about whom an SRC was asked or whom an SRC provided spontaneously.

Table 1 (see Electronic Appendix) describes 45 dataset input variables grouped into categories. Classification models were built for each of the target variables containing an answer to one of the survey questions. Table 2 (see Electronic Appendix) describes all target variables.

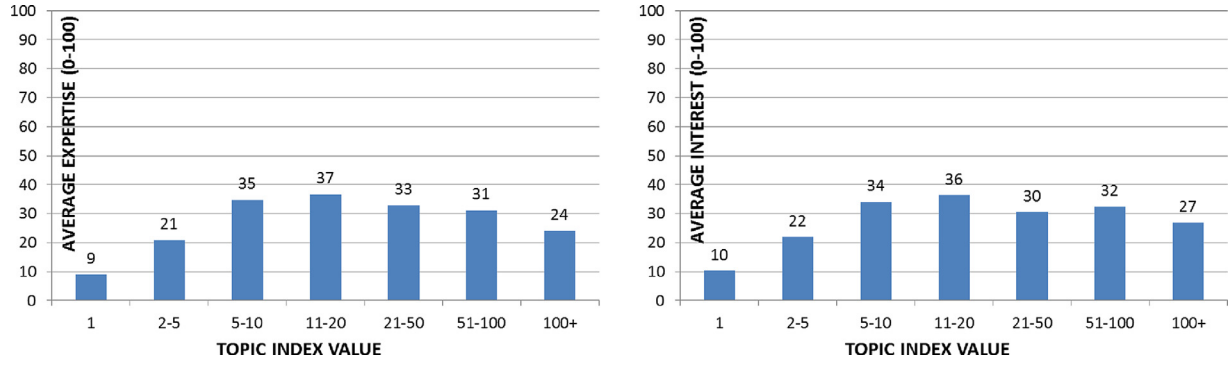

Fig. 5. Average expertise and interest value against number of edits in the topic.

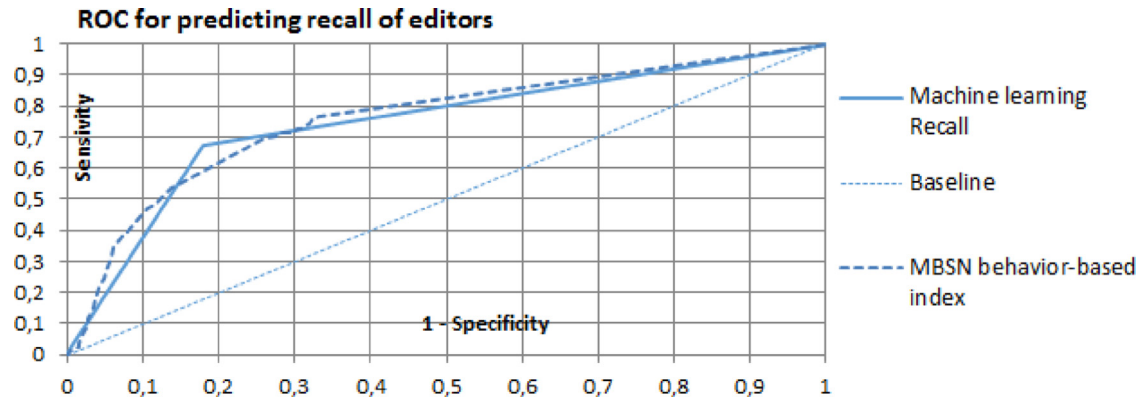

Fig. 6. ROC curve for decision tree compared to MBSN behavior-based index of acquaintance for question 1.

The described dataset contains many records with at least one missing value. Because the regression and Support Vector Machines classifiers cannot use records with missing values, we used the standard imputation methods tree surrogate (a decision tree is built for each attribute and used to fill in missing values) and median (missing values are filled in with the median) [5]. Inputting these values did not provide satisfactory results. Accordingly, we used the decision trees implemented in SAS Enterprise Miner 6.2, as they provide satisfactory handling of missing values and their decision rules are easy to interpret Fig. 5.

The trees were built with default parameter values:

- Splitting criterion based on the *p*-value of Pearson Chi-square statistic.
- Stopping criterion based on the same *p*-value and a significance level of 0.2.
- Pruning based on the 1-SE rule selecting the smallest possible tree without without significantly reducing validation accuracy.

5.2. Question 1 – spontaneous acquaintance

Q1. "We would like to know how many Wikipedians do you know. Please name every one that you can remember (use nicknames)."

We used two target class values: 1 if an *SRC* provided a *DST* s nick as an answer to the first question, 0 otherwise. The 111 surveyed users provided 292 nicks of other Wikipedians. Our dataset contains relationships for 2,668 users whom they were in a position to know, based on data extracted from Wikipedia edit history. The dataset is unbalanced: only 12.89% observations have a value of 1 for the target variable. For the machine learning process, the dataset was balanced by undersampling observations with a target value of 0 [3]. The final input dataset contains 584 observations with equal target class distribution.

After creating the decision tree, the most important variables (chosen based on the logworth value) were:

- Encounters_discussions – the number of discussions between editors *SRC* and *DST*
- Reverse_coedits – co-edits calculated for *DST* → *SRC*, edits made by *DST* near content created by *SRC*
- Co-edits – co-edits calculated for *SRC* → *DST*, edits made by *SRC* near content created by *DST*

Note that these variables make intuitive sense, as acquaintance, as defined here, need not be based solely on discussions (even though this variable is by far the most important); other interactions may contribute to recognition/recall as well.

We ran the decision tree classifier on a validation dataset (a random subset of records which were not used in the training process) to test its accuracy on future data. Moreover, the balancing procedure was not used on the test set: the original class distribution was maintained. After scoring was performed on the validation dataset, the results showed that for a balanced dataset the misclassification rate is very high, over 80%. However, when looking at the ROC curve in Fig. 6, we see a different picture. We are able to correctly identify almost 80% of all acquaintances with only a 30% false positive rate. The low accuracy stems from the fact that the negative class is much more numerous. Furthermore, Fig. 6 shows that the decision tree is superior to the behavioral index of acquaintance proposed in the previous section alone. This proves that operationalizations based on machine learning models have the potential to be far superior to single dimensions designed by hand.

Based on the computed score values, we can provide a good operationalization for Question 1. The new operationalization explains results much better than the value of the acquaintance variable. Wikipedia users remember very few other users if nicknames are not suggested.

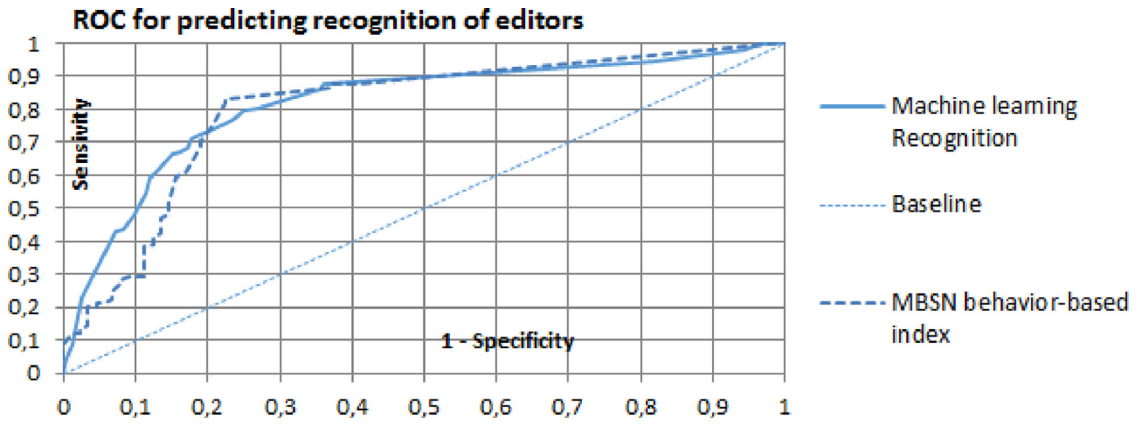

Fig. 7. ROC curve for decision tree and simple operationalization of acquaintance for question 2.

5.3. Question 2 – suggested acquaintance

In contrast to the first survey question, in the second users were presented with a list of 10 other Wikipedians whom they were in a position to know.

Q2. "Now, please look at the list of Wikipedians with whom you have edited the same articles. Mark nicknames that you recognize (you remember that you have seen them before)."

After filtering out records with DSTs not used in this question, the dataset contains 2293 observations and a slight class imbalance (61% for class 0, 39% for class 1). We undersampled in order to obtain equal target classes.

For this question, the most important variables used in decision trees were:

- encounters_discussion – number of discussions between SRC editors and DSTs.
- encounters_coedits – number of edits made by SRCs near DSTs
- d_since_l_con_discussion – number of days since the last discussion

Note that all these attributes correspond to interactions which, intuitively, may contribute to higher recall.

From the rather complex decision tree we manually selected the two simplest rules which could help in determining the probability of recognition:

1. If $encounters\_discussion \geq 3$ then we have a 77% probability that SRC chooses a DST from the proposed list.
2. If $encounters\_discussion = 0$ and $encounters\_coedits \leq 8$ then we have an 88% probability that SRC will not choose a DST.

In order to evaluate the newly-created decision tree, we generated ROC curves for it and for the earlier operationalization (acquaintance). Fig. 7 shows that the result from the decision tree is far superior to that from the acquaintance variable only, and that the synthetic index can be viewed as approximately equivalent to its social counterpart.

5.4. Question 4 – high quality edits

Let us now try to operationalize the notion of trust based on:

Q4. "Please mark the nicks of editors that have, in your opinion, edits of a good quality."

The model was built based on 1116 observations. The most important variables were:

- src_dst_disc_210_bool – binary indicator if the discussion variable for SRC editors and DSTs is higher than 210.
- dst_total_edits – number of edits made by DSTs.
- encounters_discussion – number of discussions between SRC editors and DSTs.

In Fig. 8 we present the ROC curve comparing regression for the Trust operationalization described earlier and the decision tree. The decision tree shows a great improvement, although we obtained an even better fit for the first two questions.

An analysis of the variables used by the model gives an insight into the correlation between the Discussion network and responses to Question 4 (declared trust in an editors competence). The model used the Discussion dimension (*dst_src_disc*), indicating that an editor is more aware that other editors comment on talk pages than that they are making edits next to the work of other editors. This has a common-sense explanation: when editors make edits of their own, the user interface does not show the authorship of nearby edits. On the other hand, when editors enter the talk pages they become aware of other editors entries, and can learn in the discussion history who the authors of these entries are. As it turns out, this is more likely to create an awareness of edits made by another editor.

The model also used the encounters_discussion variable (derived, like the Discussion network, from the Wikipedia talk pages, but indicating only the number of times that two editors exchanged any information on the talk pages, without

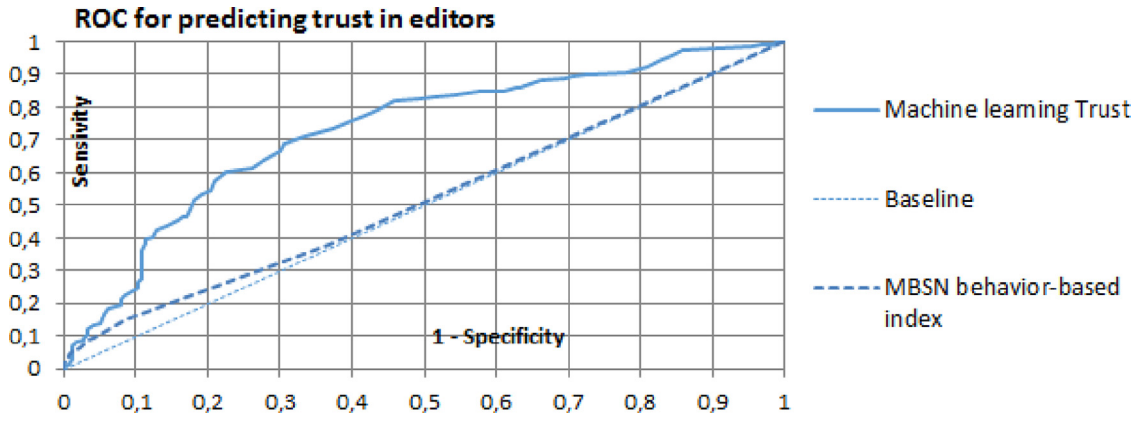

Fig. 8. ROC curve for decision tree and regression of trust and new operationalization for question 4.

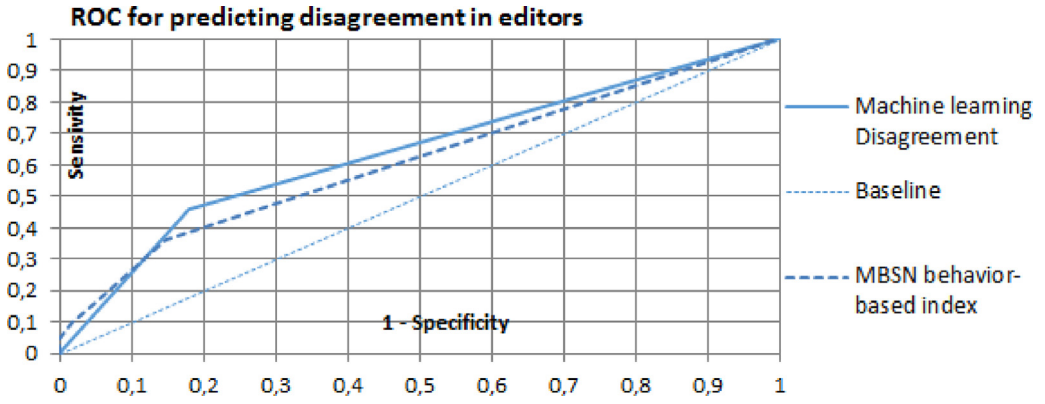

Fig. 9. ROC curve for decision tree and regression of disagreement for question 5.

giving the length of the discussion); nevertheless, it chose a much higher cutoff point than in the case of the first two questions regarding spontaneous and supported editor acquaintance. For Questions 1 and 2, the minimum strength of the encounters_discussion variable was 4 and 3, respectively, while for Question 4 the cutoff point was 14. This indicates that a prolonged discussion on the talk pages is likely to create trust among editors.

Since the model did not use the Co-edits network, our improved fit of responses to Question 4 does not change the conclusion about the validity of Hypothesis 2. Interestingly, the model also used a simple count of the number of edits made by the evaluated editor. This suggests that the number of an editors edits are tied to his/her trustworthiness.

5.5. Question 5 – disagreement or argument

For this question we proposed a list of nicks with whom respondents might have argued. The list was generated based on revert history.

Q5. "Please select the nicks of editors you have at any time disagreed with or argued with." After filtering out records with DSTs not used in this question, we were left with 1116 observations. The data contained only 18% confirmed arguments; moreover, for the creation of a decision tree, we undersampled the dataset to create a balanced dataset with 400 observations. We created a tree with a misclassification rate of 24%, with the three most important variables being:

- d_since_l_con_revert – number of days since the last revert
- d_since_f_con_coedit – number of days since the first coedit
- dst_cat_div_20 – number of categories in which a DST editor contributed at least 20 edits

The decision tree for Question 5 is extremely complex. Scoring performed on 547 test cases showed improvements over previous operationalizations, but not sufficient to declare equivalence. The respective ROC curves are given in Fig. 9.

We defined a better model for Question 5 based on computed variables. The resulting decision tree uses the Reverts dimension along with the Co-edits dimension. This can be interpreted as follows: an absence of reverts together with a sufficiently high number of co-edits indicates that two editors worked together, were aware of each others edits, and are not likely to be in conflict. However, the overall obtained classification accuracy is too low to conclude that Hypothesis 3 is valid.

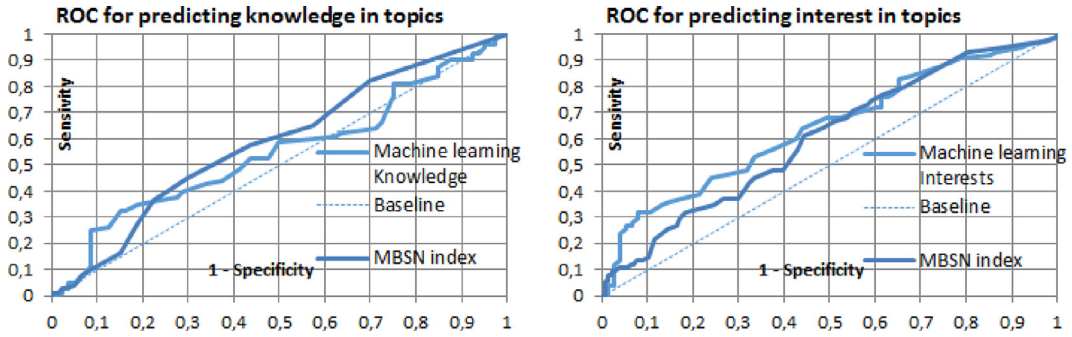

Fig. 10. Predicting knowledge and interest of editors in specific topics.

5.6. Knowledge and interests

To analyze the last two questions, we prepared a separate dataset, since these questions are based not on interactions between two users but on user activity in particular categories. Users were asked to answer questions about their interest in a given topic and their knowledge in this field:

Q7. "Please look at the list of article categories. For each category try to determine your level of competency". [Scale] 0 = "I don't know much about it" 100 = "I know a lot about it".

Q8. "And now for each category please determine how much you are interested in that topic". [Scale] 0 = "I'm not interested at all" 100 = "I'm very interested".

Responses given by users were binarized (50 or higher was considered class 1, otherwise class 0). The dataset contained 1,092 observations for 104 different Wikipedians; the target classes are unbalanced (class 1 = 76%; class 0 = 24%).

To test which variables are useful for determining a Wikipedians interest or knowledge in a given category, we created a series of prediction models. We used standard imputation methods to fill in missing values. To correct the problem with an imbalanced target variable, we undersampled the dataset.

We were unable to build a suitable decision tree for knowledge or interest. Instead, we used gradient boosting and logistic regression with default parameters. For Question 7 we were unable to create a model better than the respective value from the MBSN. For Question 8, variables important for predicting the answer were:

- no_rev_2010 – number of reverts in a topic in 2010 year
- val_coedits_2011 – length of edits in a topic in 2011 year
- no_coedits_2011 – number of edits in a topic in 2011 year
- no_rev_2010 – number of reverts in a topic in 2010 year

The misclassification rate for the regression model for Question 7 was 43%; for Question 8, 40%. Fig. 10 presents two ROC curves for models using only MBSN and other variables. The models predictions were quite inaccurate, leading to the conclusion that the interests and expert knowledge of surveyed editors cannot be accurately predicted based on available behavioral data.

6. Conclusions and future work

In this paper, we have aimed to check the validity of a set of hypotheses regarding social interpretations of behavioral social networks created from the Wikipedia edit history. Undoubtedly, the subject of modeling the Wikipedia knowledge community using social networks deserves the attention of scientists and practitioners interested in improving open collaboration or knowledge management used by Wiki or other Web 2.0 technology. Wikipedia is unique because it is one of the largest, oldest, and best-known knowledge communities, and also because of the availability of a detailed edit history that records almost all behavior on the part of editors. The hypotheses we formulated regarding social interpretations were largely based on related work in the area, as well as common sense. For these reasons, it came as a surprise that only one of them (Hypothesis 1. *The Discussion network can be interpreted as acquaintance among editors*) turned out to be valid. Several other hypotheses, such as the interpretation of the Reverts network as conflict or of the Topics network as interest or expertise, seemed intuitive and were supported by indirect evidence in previous work. The lack of validity of these hypotheses may be the result of too simple a definition of the related behavioral networks; however, our subsequent use of the data mining approach revealed a much deeper problem. In the latter stage of our research, we attempted to create improved operationalization of behavioral social networks that could best classify survey data. The goal of this approach was superior verification of the hypotheses, as the initial definition of the MBSN was very simple. We used all available variables, including some related to time (based on the hypothesis that short-term human memory is the factor that accounts for the inaccuracy of predicting declarative networks using behavioral ones). While this approach revealed a few useful rules, and

an improved model for predicting survey data related to edits of high quality, this model used not the Co-edits dimension postulated by Hypothesis 2 but the Discussion network and the total number of edits by the evaluated editor.

Our models had only a limited ability to predict editor conflict. We were completely unable to predict user responses to questions about interests or expertise based on available behavioral data

These results show that it is much harder than expected to model real social phenomena, even based on a complete and detailed history of behavior. Of course, our study had several limitations that can partially explain these difficulties. The main one was the lack of a random sample of Wikipedia editors, and, in particular, an overrepresentation of active editors and administrators. This may have led to difficulties in predicting interests and expert knowledge, since admins and candidates for admins on Wikipedia edit massive numbers of articles, regardless of their interests and knowledge, in search of minor improvements and in order to implement Wikipedia editing policies. However, even when we looked at only larger edits (to exclude minor improvements), we failed to obtain a behavioral network with a greater capacity to predict survey data.

In our view, a limitation much more important than the sample composition may account for the difficulty in using behavioral data to predict survey data. On Wikipedia, edits are often anonymous, and the Wiki user interface itself does not support the easy determination of authorship of the various parts of an articles text. This makes it hard for editors to know who co-edits with them or reverts their edits. Of course, in some cases, edit wars occur or editors do become aware of the quality of their co-editors. However, such occurrences are relatively rare, and therefore the disagreement between behavioral and declarative networks is great.

These observations do not preclude the use of behavioral social networks for practical applications, such as recommending good candidates for admins, or for testing hypotheses regarding open collaboration (for example, regarding the impact of coordination on talk pages on article quality). Social hypotheses concerning influence or social capital may be especially easy to test using the validated Discussion network. However, one should be wary of applications that rely on a direct social interpretation of a behavioral network, for example on the interpretation of Reverts as conflict or Topics as interest or expert knowledge. Such applications should rely on independent tests of the impact of using behavioral networks on application functionality.

Acknowledgments

The research is partially supported by the Polish National Science Centre grant 2012/05/B/ST6/03364 and the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 316097 [ENGINE].

References

[1] E.R. Babbie, The Practice of Social Research, Wadsworth Publishing, Cengage Learning, Belmont, 2010.
[2] P. Borzymek, M. Sydow, A. Wierzbicki, Enriching trust prediction model in social network with user rating similarity, in: Proceedings of the CASON'09. International Conference on Computational Aspects of Social Networks, IEEE, 2009, pp. 40–47.
[3] N. Chawla, Data mining for imbalanced datasets: An overview, in: O. Maimon, L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook, Springer US, 2010.
[4] N. Criado, J.M. Such, Implicit contextual integrity in online social networks, Inf. Sci. 325 (2015) 48–69.
[5] A. Farhangfar, L. Kurgan, J. Dy, Impact of imputation of missing values on classification error for discrete data, Pattern Recognit. 41 (12) (2008) 3692–3705.
[6] T. Fawcett, An introduction to roc analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.
[7] Y. Gandica, F.S. dos Aidos, J. Carvalho, The dynamic nature of conflict in wikipedia, EPL (Europhys. Lett.) 108 (1) (2014) 18003.
[8] D. Gergle, N. Contractor, B. Keegan, Do editors or articles drive collaboration?: multilevel statistical network analysis of wikipedia coauthorship., in: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW'2012), ACM, New York, 2012, pp. 427–436.
[9] M. Jankowski-Lorek, L. Ostrowski, P. Turek, A. Wierzbicki, Wikipedia knowledge community modeling, in: R. Alhajj, J. Rokne (Eds.), Proceedings of the Encyclopedia of Social Network Analysis and Mining, Springer New York, 2014, pp. 2410–2420.
[10] T. Kaszuba, A. Hupa, A. Wierzbicki, Advanced feedback management for internet auction reputation systems, IEEE Internet Comput. 14 (5) (2010) 31–37.
[11] P. Kazienko, K. Musial, E. Kukla, T. Kajdanowicz, P. Brdka, Multidimensional social network: model and analysis, in: ICCCI'11 Proceedings of the Third International Conference on Computational Collective Intelligence: Technologies and Applications - Volume Part I, ICCI, 2011, pp. 378–387.
[12] A. Kittur, B. Suh, B.A. Pendleton, E.H. Chi, He says, she says: Conflict and coordination in wikipedia, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2007, ACM, New York, 2007, pp. 453–462.
[13] R.K.A. Kittur, Beyond wikipedia: coordination and conflict in online production groups., in: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW'2010), ACM, New York, 2010, pp. 215–224.
[14] M.-T. Le, H.-V. Dang, E.-P. Lim, A. Datta, Wikinetviz: Visualizing friends and adversaries in implicit social networks, in: Proceedings of the ISI, 2008, pp. 52–57.
[15] N. Leibowitz, M. Ripeanu, A. Wierzbicki, Deconstructing the kazaa network, in: WIAPP 2003. Proceedings of the Third IEEE Workshop on Internet Applications, IEEE, 2003, pp. 112–120.
[16] P. Levy, Collective Intelligence, Perseus Books, 1997.
[17] X. Li, J. Tang, T. Wang, Z. Luo, M. de Rijke, Automatically assessing wikipedia article quality by exploiting article editor networks, in: A. Hanbury, G. Kazai, A. Rauber, N. Fuhr (Eds.), Advances in Information Retrieval, *Lecture Notes in Computer Science*, vol. 9022, Springer International Publishing, 2015, pp. 574–580.
[18] M. Morzy, A. Wierzbicki, The sound of silence: mining implicit feedbacks to compute reputation, in: Proceedings of the Internet and Network Economics, Springer, 2006, pp. 365–376.
[19] R. Nielek, A. Wawer, A. Wierzbicki, Spiral of hatred: social effects in internet auctions. between informativity and emotion, Electron. Commer. Res. 10 (3–4) (2010) 313–330.
[20] L. Ostrowski, P. Turek, A. Wierzbicki, M. Jankowski-Lorek, Modeling wikipedia admin elections using multidimensional behavioral social networks 3 (4) (2013) 787–801.

[21] A. Passant, J. Breslin, J. Schneider, A qualitative and quantitative analysis of how wikipedia talk pages are used, in: Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, ACM, Raleigh, NC, 2010.
[22] R. Priedhorsky, J. Chen, S.T.K. Lam, K. Panciera, L. Terveen, J. Riedl, Creating, destroying, and restoring value in wikipedia, in: Proceedings of the 2007 International ACM Conference on Supporting Group Work, GROUP 2007., ACM, New York, 2007, pp. 259–268.
[23] K. Rzadca, D. Trystram, A. Wierzbicki, Fair game-theoretic resource management in dedicated grids, in: Proceedings of the CCGRID 2007, Seventh IEEE International Symposium on Cluster Computing and the Grid, IEEE, 2007, pp. 343–350.
[24] D. Spinellis, P. Louridas, The collaborative organization of knowledge, in: Proceedings of the Communications of the ACM - Designing Games with a Purpose, vol. 51 (8), 2008.
[25] B. Suh, G. Convertino, E.H. Chi, P. Pirolli, The singularity is not near: Slowing growth of wikipedia, in: Proceedings of the 5th International Symposium on Wikis and Open Collaboration (WikiSym 2009), Orlando, Florida, 2009, pp. 1–10.
[26] P. Turek, A. Wierzbicki, R. Nielek, A. Hupa, A. Datta, Learning about the quality of teamwork from wikiteams, in: Proceedings of the 2010 IEEE second international conference on social computing, SocialCom/IEEE international conference on privacy, security, risk and trust, PASSAT 2010., Minneapolis, 2010, pp. 17–24.
[27] P. Turek, J. Spychaa, A. Wierzbicki, P. Gackowski, Social mechanism of granting trust basing on polish wikipedia requests for adminship, in: Proceedings of the international conference on social informatics (SocInfo 2011), Singapore, 2011, pp. 212–225.
[28] P. Turek, A. Wierzbicki, R. Nielek, A. Hupa, A. Datta, Wikiteams: How do they achieve success? IEEE Potentials 30(5) (2011) 2–7.
[29] F.B. Viégas, M. Wattenberg, K. Dave, Studying cooperation and conflict between authors, in: Proceedings of the 2004 conference on human factors in computing systems, ACM, New York, 2004.
[30] B.-Q. Vuong, E.-P. Lim, A. Sun, M.-T. Le, H.W. Lauw, K. Chang, On ranking controversies in wikipedia: Models and, WSDM 2008 (2008) 171–182.
[31] A.D. Williams, D. Tapscott, Wikinomics, Penguin Books, 2008.
[32] Y. Zhang, A. Sun, A. Datta, K. Chang, E.-P. Lim, Do wikipedians follow domain experts?, A Domain-specific Study on Wikipedia Knowledge Building, JCDL 2010, 2010, pp. 119–128.